GAM Practice

STAT 245

Table of contents

Data 1
What to Do And Hand In 3
Tasks are listed in order of priority. 3
Get as far as you can in the time you have. 3
Save your work as an qmd file, turn in a rendered copy as a team at the end of the work session. 3
About Random Effects in GAMs 3
Another Example **If you have extra time ONLY** 4
More Data Options 5

Data

Before you dive in: there are alternative options for datasets to work with at the bottom of this instruction document. You can choose!

The code below reads in a dataset with public health data from Zambia and does some data cleaning. The response variable of interest is `height_zscore`, the z-score of the child's height compared to the national average at their age. (The assumption is that malnourished or unhealthy children will be unusually small, and many of these children are, especially at older ages.) Other variables include:

- `child_sex`
- `breastf` duration of breast-feeding in months
- `child_age` child's age in months
- `mother_birth_age` mother's age when the child was born, in years
- `mother_height` mother's height in cm
- `mother_BMI` mother's body mass index
- `mother_education` mother's education level
- `mother_work` mother's work status

- `region` Region in Zambia of mother's residence
- `district` District in Zambia of mother's residence

```
zam <- read.table('https://raw.githubusercontent.com/cran/sdPrior/master/inst/examples/zam
                  header=TRUE)
names(zam) <- c('height_zscore', 'child_sex', 'breastf', 'child_age',
                'mother_birth_age', 'mother_height', 'mother_BMI',
                'mother_education', 'mother_work', 'district', 'region', 'time')
zam <- zam |>
  mutate(child_sex = ifelse(child_sex == 1, 'Male', 'Female'),
         mother_education = factor(mother_education),
         mother_education = fct_recode(mother_education,
                                       'None' = '1',
                                       'Primary School' = '2',
                                       'Secondary School' = '3',
                                       'Higher Education' = '4'),
         mother_work = ifelse(mother_work==1, 'Working', 'Not Working'),
         region = factor(region),
         region = fct_recode(region,
                             'Central' = '1',
                             'Copperbelt' = '2',
                             'Eastern' = '3',
                             'Luapula' = '4',
                             'Lusaka' = '5',
                             'Northern' = '6',
                             'Northwestern' = '7',
                             'Southern' = '8',
                             'Western' = '9'),
         district = factor(district)) |>
  dplyr::select(-time)
zam <- arrange(zam, district)
glimpse(zam)
```

```
Rows: 4,421
Columns: 11
$ height_zscore    <int> -264, -389, -127, -169, -156, -269, -169, 5, -279, 10~
$ child_sex        <chr> "Male", "Female", "Female", "Male", "Male", "Female",~
$ breastf          <int> 24, 19, 1, 24, 0, 0, 16, 14, 19, 11, 1, 40, 21, 21, 0~
$ child_age        <int> 29, 57, 16, 46, 9, 5, 30, 56, 25, 13, 16, 46, 32, 33,~
$ mother_birth_age <dbl> 25.58333, 23.25000, 35.66667, 33.16667, 31.25000, 35.~
$ mother_height    <dbl> 162.4, 162.4, 151.8, 151.8, 156.6, 161.1, 161.1, 161.~
$ mother_BMI       <dbl> 22.33, 22.33, 18.66, 18.66, 24.22, 25.58, 25.58, 25.5~
```

```
$ mother_education <fct> Primary School, Primary School, Primary School, Prima~
$ mother_work      <chr> "Working", "Working", "Working", "Working", "Working"~
$ district         <fct> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1~
$ region           <fct> Northern, Northern, Northern, Northern, Northern, Nor~
```

What to Do And Hand In

Tasks are listed in order of priority.

Get as far as you can in the time you have.

Save your work as an qmd file, turn in a rendered copy as a team at the end of the work session.

1. Using the Zambia dataset, explore the data a little, then fit your a regression model with `height_zscore` as the response variable and predictors chosen from among: `mother_height`, `mother_birth_age`, `breastf`, and `mother_education`.

- *Model planning should be very rapid to save time, but do discuss which predictor(s) should be included as smooths, and of course make sure your model includes at least one smooth term.*

- For each smooth, make a note to explain: *How did you decide which predictors required smooth terms? What smoothing basis (`bs = ...`) will you use? What basis dimension `k` did you choose and why?*
- Interactions between smooth and other terms require special coding syntax in the model formula, so omit those for this first try.

2. View the model summary and, more importantly, make prediction plots - what do you learn from them? If time, you can do model selection to aid interpretation.
3. Carry out model assessment. (We should have done this before interpretation! But, in order to prioritize learning how to interpret GAMs and knowing your time was limited, this got 3rd priority…) If you find a problem with the ACF, consider adding a random effect of `district` or `region` (but recall we can't nest them in a simple `gam()`).

About Random Effects in GAMs

- Is there any need for random effects? (fitting a GAMM (GAM with not-nested random effects) is a bit beyond our course goals, but you can give it a try as directed below). Basically, we add a smooth to our `gam()` model with syntax `s(RE_variable, bs = 're')`.

```
mymodel <- gam(response ~ predictor1 + s(predictor2, ...) +
                 ... + predictorN +
                 s(random_effect_variable, bs = 're'),
                 family = ... , data= ..., method = 'ML')
```

Another option (slightly more effort and may fit more slowly) is to use the function `gamm4()` *from the package* `gamm4` *to fit the model. The syntax is then similar to* `(g)lmer()`*. You may want to try this method if you need nested random effects.*

```
library(gamm4)
mymodel <- gamm4(response ~ predictor1 + s(predictor2, ...) +
                 ... + predictorN +
                 random = ~(1|random_effect_variable),
                 family = ... , data= ..., REML=...)
# to view summary...
summary(mymodel$mer)
# to get residuals...
resid(mymodel$mer)
# generally you will work with the mymodel$mer output object and not the mymodel$gam.
```

Another Example If you have extra time ONLY

(Probably not a GAM problem - this is just practice with REs.) Fit a model for the problem below and do model assessment for your fitted model, correcting as needed for any problems found.

The Panel Study of Income Dynamics (PSID) is a longitudinal study of a representative sample of US individual. The study started in 1968 and has tracked people over time, recording data on their `age`, `educaction`, `sex`, `income`, and a personal ID number (`person`). The data are stored in dataset `psid` in package `faraway` (run the line, `library(faraway)` to load the data - there's nothing else needed to "read it in). It is suggested that you fit a model for log(`income`) as your response, and if you use the `year` predictor, first center it by subtracting the median value. The code below creates the variables `log_income` and `scaled_year` for you. Things to consider as you set up your model (you don't have to include written answers to these questions, just think before you fit): *What predictors would you include – any interactions? What random effect(s) or grouping variables would you use? Would you consider a random slope model for these data?*

```
library(faraway)
psid <- psid |>
  mutate(log_income = log(income),
```

```
        scaled_year = year - median(~year, data = psid, na.rm = TRUE))
```

More Data Options

Still curious, or don't want to try the Zambia child height data? Here are more dataset options for GAM exploration:

Note, To read in data files that are in tab-delimited text format, you can use the function `readr::read_table()`*:*

- How does pitch frequency change over time as a person sings? http://www.statsci.org/data/general/ooh.html
- How does the magnitude of light emitted by star RR Lyrae 1263 vary over time? http://www.statsci.org/data/oz/rrl1263.html
- Can you predict the rainfall in Canberra? http://www.statsci.org/data/oz/wind_ca.html
- How do PCB concentrations in trout change as trout age? http://www.statsci.org/data/general/troutpcb.html
- How have the winning times in the Sydney to Hobart Yacht race changed over time? http://www.statsci.org/data/oz/sydhob.html
- What factors affect the number of species (your choice of group - finches, plants, endemic finches or plants…) present in the Galapagos? http://www.statsci.org/data/general/galapagos.txt
- What factors affect acidity of rainfall water in the UK? http://www.statsci.org/data/general/rainuk.html
- Consider data from the Scripps Institution of Oceanography CO_2 Program. Model the total dissolved inorganic carbon (`DIC`) in sea water over time at three observation `Station`s, and as a function of `Depth`,, `Salinity`, and `Temp`erature.
- Consider a dataset giving details on `price` and other characteristics of 6172 Lego sets, including the number of `Pieces` and `Minifigures` inside and the `Year` they were created, as well as the `Name`, `Theme`, `Subtheme` and `Item_Number` of each set. Data are at: https://sldr.netlify.app/data/legos.csv. Try to predict the price of a lego set.